The Path to Malaria Elimination & Techniques to Detect Offensive Language

Filipa Peleja

filipa.peleja@vodafone.com

28 March 2019

Data Science Research at Vodafone

≈ 125 Data scientists with PhDs and MScs from top international universities

25 Academic collaborations

17 Research papers published in the past 2 years

81 Conferences attended, reaching more than 12K attendees

10 Data Science event sponsorships

21 International awards and recognitions since 2016

C1

icons from https://www.flaticon.com

Data Science Research for Social Good

4.9+ billion mobile phone subscribers worldwide

66% of worlds' population

Mobile penetration of 120% to 89% of population

More time spent on our phones than watching TV or with our partner (US and UK)

Emerging and developed regions

Data Science Research for Social Good

4.9+ billion mobile phone subscribers worldwide

66% of worlds' population

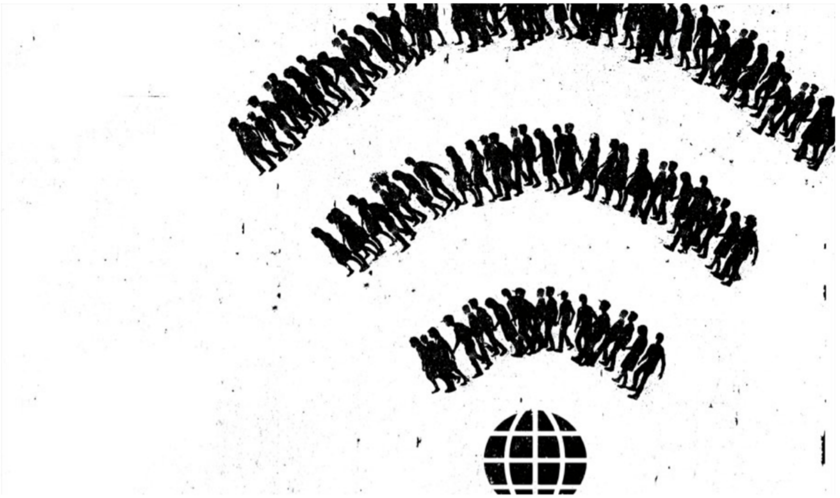
Mobile penetration of 120% to 89% of population

More time spent on our phones than watching TV or with our partner
(US and UK)

Emerging and developed regions

Big Data from Cheap Phones

10 Breakthrough Technologies The List + Years +



Big Data from Cheap Phones

Collecting and analyzing information from simple cell phones can provide surprising insights into how people move about and behave—and even help us understand the spread of diseases.

C1

https://www.technologyreview.com/s/513721/big-data-from-cheap-phones/

This map, a product of cell-phone data analytics, shows the most important sources of malaria infections (darker shades)—taking into account the potential for further transmission caused by human travel—as well as the major destinations of people exposed to the disease (lighter shades). It can be used to determine where best to focus warnings and mosquito control techniques.

"This is the future of epidemiology. If we are to eradicate malaria, this is how we will do it."

"We can really provide not just insight, but actually something that is actionable. This really does work."

Call Detail Records: Typical Mobile Data

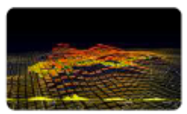
VOICE

HR_ORG	TLFN_A	TLFN_B	CD_GEO_A	CD_GEO_B	DT_ORG	CD_SNTD	CD_ERB	CD_CCC	QT_DUR
20:05:31	XXX	YYY	3	11	20140519	2	1562	568	33
...	...	...	...	...	...	...	...	...	...

SMS

HR_ORG	TLFN_A	TLFN_B	CD_GEO_A	CD_GEO_B	DT_ORG	CD_SNTD	QT_TRFG
15:53:54	XXX	ZZZ	3	25	20140506	2	1
...	...	...	...	...	...	...	...

Consumption	Social Network	Mobility
Call duration	In/Out Degree	Radius of gyration
N. Events	Delta w.r.t time window	Travelled distance
Lapse between events	Unique Calls per day	Rate of popular antennas
Reciprocated events	Unique SMS per day	Regularity of popular antennas



C1

Mobile Network Data has clear Advantages

Cost and Effort	• Most of the mobile data that could be used for the public sector is data that has been collected already for other purposes. In addition, mobile data is typically collected by automatic means which makes its collection very cost-efficient
Temporal and Spatial Granularity	• Mobile data can be available in real-time or if not real time much more frequently than how data is typically collected (every 5-10 years for census data) • Some types of mobile data can be collected with significantly finer grained spatial granularity than with traditional methods
Accuracy and Scale	• It could be argued that some kinds of data that are relevant for the public sector (e.g. migrations) can be collected more accurately by automatic means than by manual means as it is the state-of-the-art • In addition, given that there isn't a human-in-the-loop, the data is less prone to human errors and potential biases introduced by humans

Mobile data brings value for public good and positive social impact in variety of areas

Joint work from Pedro Rente Lourenço, Dr. Nuria Oliver, Jessica Floyd, Prof. Andy Tatem and Dr. Nick Ruktanonchai

Epidemiological Studies
The path to Malaria Elimination

Source: The path to eradication: a progress report on the malaria-eliminating countries, Gretchen Newby, Adam Bennett, Erika Larson, Chris Cotter, Rima Shretta, Allison A Philips, Richard G A Feachem, The Lancet, 387, 10029, pp 1775-1784, 2016

Mozambique

- Malaria accounts for over half of all outpatients visits and over a quarter of all hospitalizations

- Mobility is a key factor threatening success

- The analysis of aggregate mobile data could provide a route to significant and rapid impact

C1

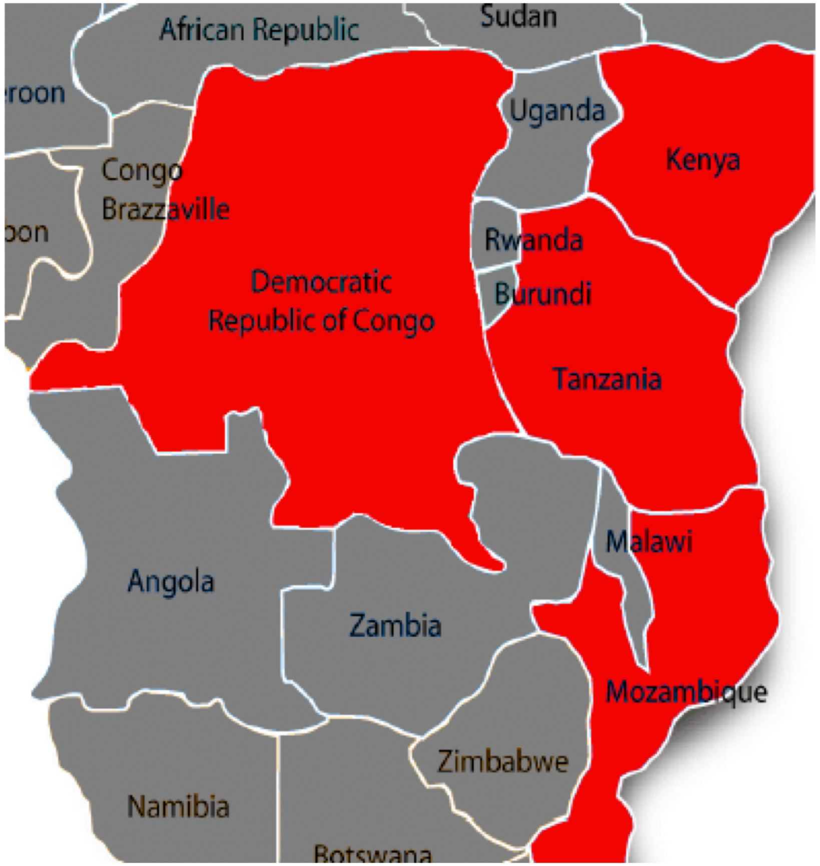
Joint work from Pedro Rente Lourenço, Dr. Nuria Oliver, Jessica Floyd, Prof. Andy Tatem and Dr. Nick Ruktanonchai

Thinking regionally

- Parasites do not respect national borders

- Regional analysis have substantial benefits to wider continental efforts to tackle the disease

Through the analysis of aggregated, pseudonymized CDRs, we can help in malaria elimination



C1

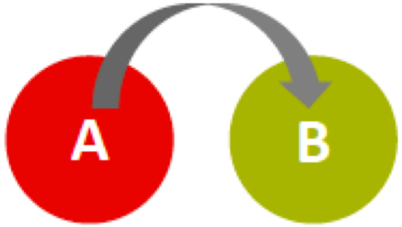
Joint work from Pedro Rente Lourenço, Dr. Nuria Oliver, Jessica Floyd, Prof. Andy Tatem and Dr. Nick Ruktanonchai

Data sources in VDF research

- **Mobile data**: Pseudonymized CDRs from February to May 2018
- **Population data**: WorldPop population density estimates
- **Malaria data**: Monthly malaria incidence from February to May of 2017

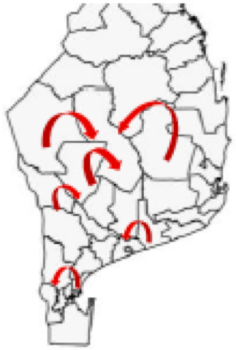
Joint work from Pedro Rente Lourenço, Dr. Nuria Oliver, Jessica Floyd, Prof. Andy Tatem and Dr. Nick Ruktanonchai

Malaria mobility

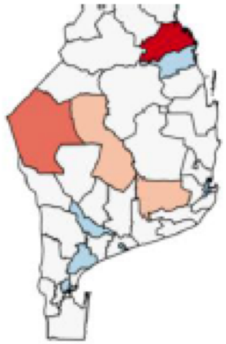


Mobility matrices
to see human movement

Malaria Incidence
scaled by population per district



Malaria Mobility
as a result of the previous two



Malaria Sinks and Sources
by scaling at population level

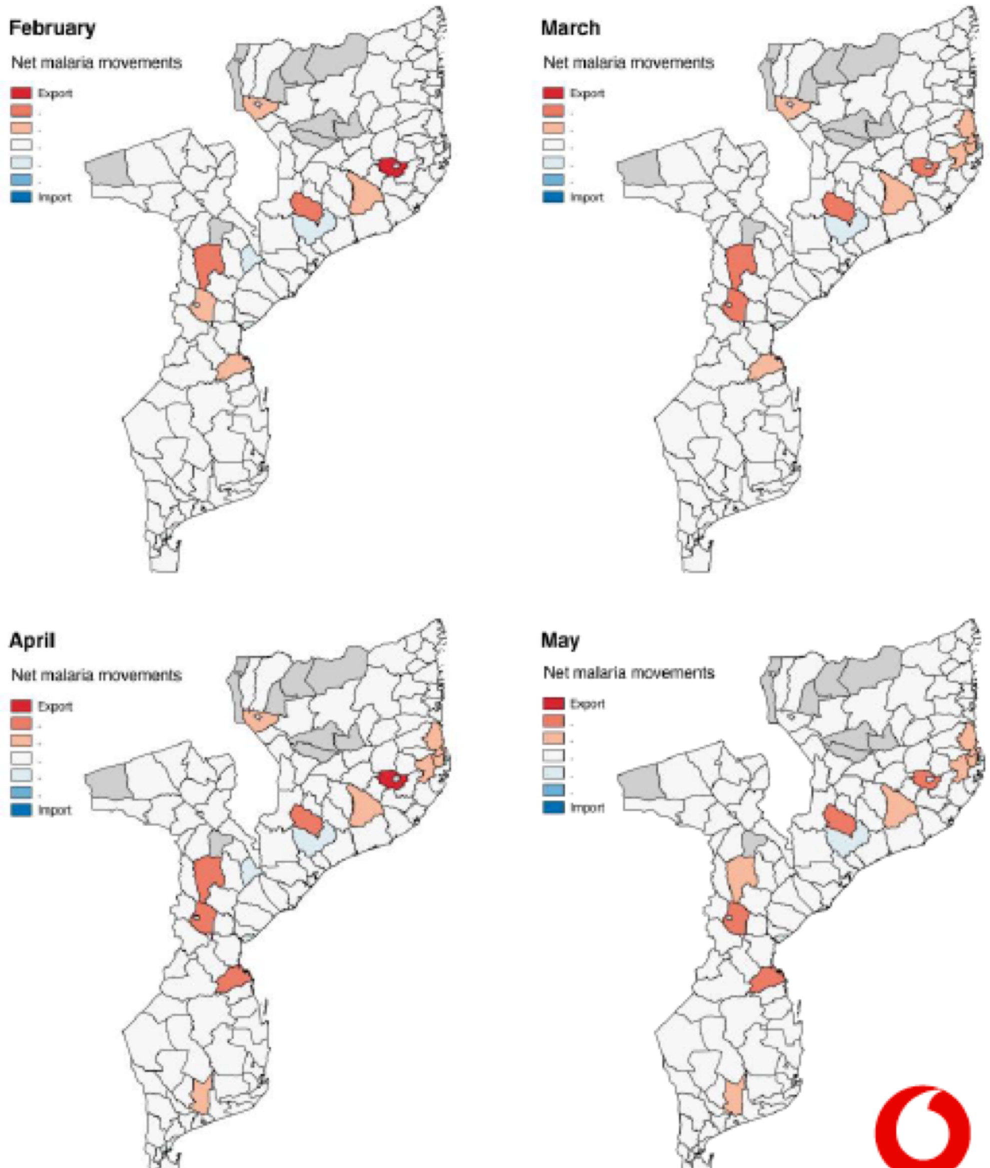
Joint work from Pedro Rente Lourenço, Dr. Nuria Oliver, Jessica Floyd, Prof. Andy Tatem and Dr. Nick Ruktanonchai

Sinks and sources of malaria

- The maps illustrate the seasonality of malaria
- Districts in grey did not have enough data to make reliable estimates

- These maps show the relative net number of movements with malaria by district

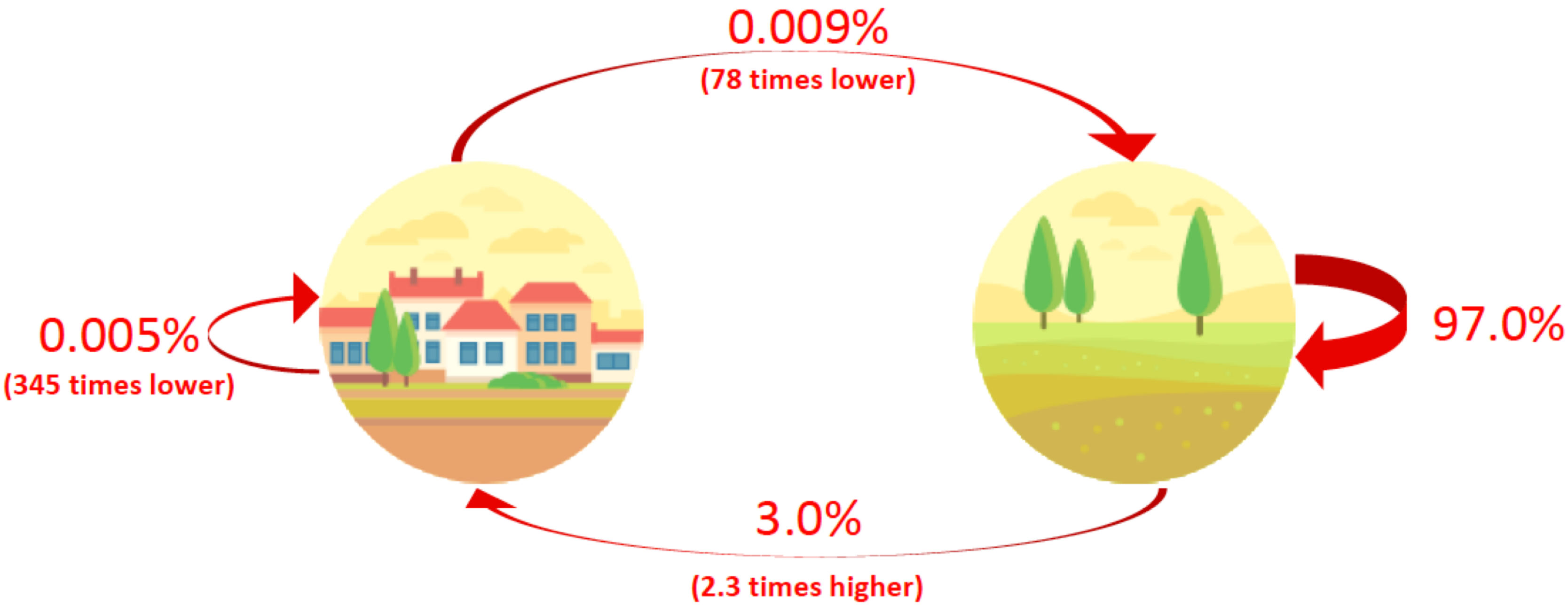
- **Small, densely populated districts tended to have more malaria imported than exported while larger, more rural districts tend to export more**



C1

Joint work from Pedro Rente Lourenço, Dr. Nuria Oliver, Jessica Floyd, Prof. Andy Tatem and Dr. Nick Ruktanonchai

Urban and Rural movements

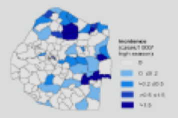


Q: Is there a significant difference in importing/exporting malaria between rural and urban areas?

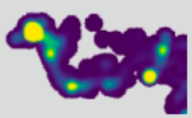
A: Yes! Rural areas export more malaria to urban areas than the other way around

Joint work from Pedro Rente Lourenço, Dr. Nuria Oliver, Jessica Floyd, Prof. Andy Tatem and Dr. Nick Ruktanonchai

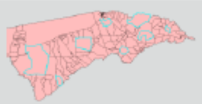
Where Vodafone can help at the Four challenges in malaria elimination



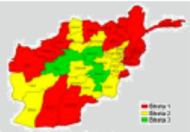
Measuring incident



Mapping Transmission focus of infection



Testing interventions



Elimination strategy design

Joint work from Pedro Rente Lourenço, Dr. Nuria Oliver, Jessica Floyd, Prof. Andy Tatem and Dr. Nick Ruktanonchai

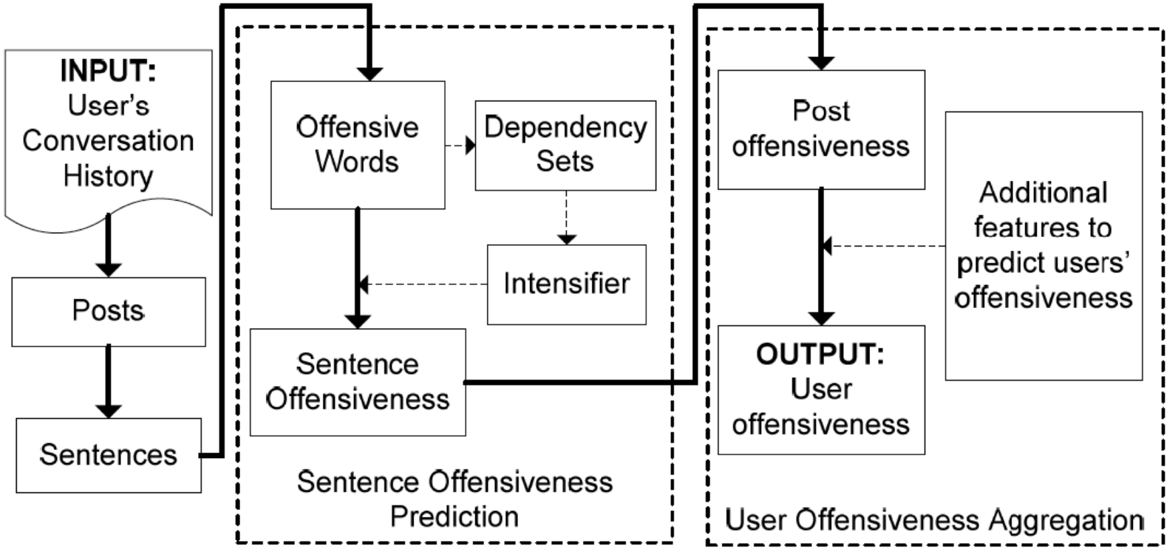
Techniques to Detect Offensive Language

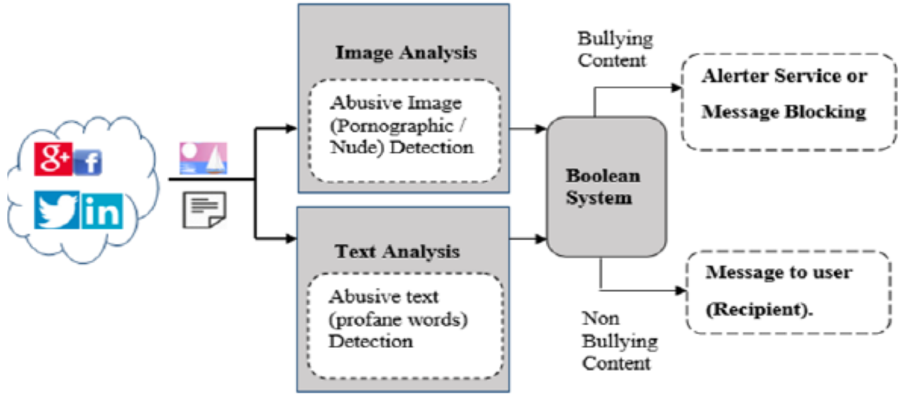
Joint work from Filipa Peleja and Sara Hajian

Offensive and discriminatory language

- People communicate online more and more. Social networks play an important role but not only: blogs, official and non-official news feeds, etc. disseminate information about various topics
- In this environment many times we witness improper behavior used to intimidate and undermine individuals, and with the popularity of these tools such behavior has been growing rapidly
- This behavior can lead to hazardous outcomes in terms of impact it has on social networks

Detection of offensive and discriminatory language





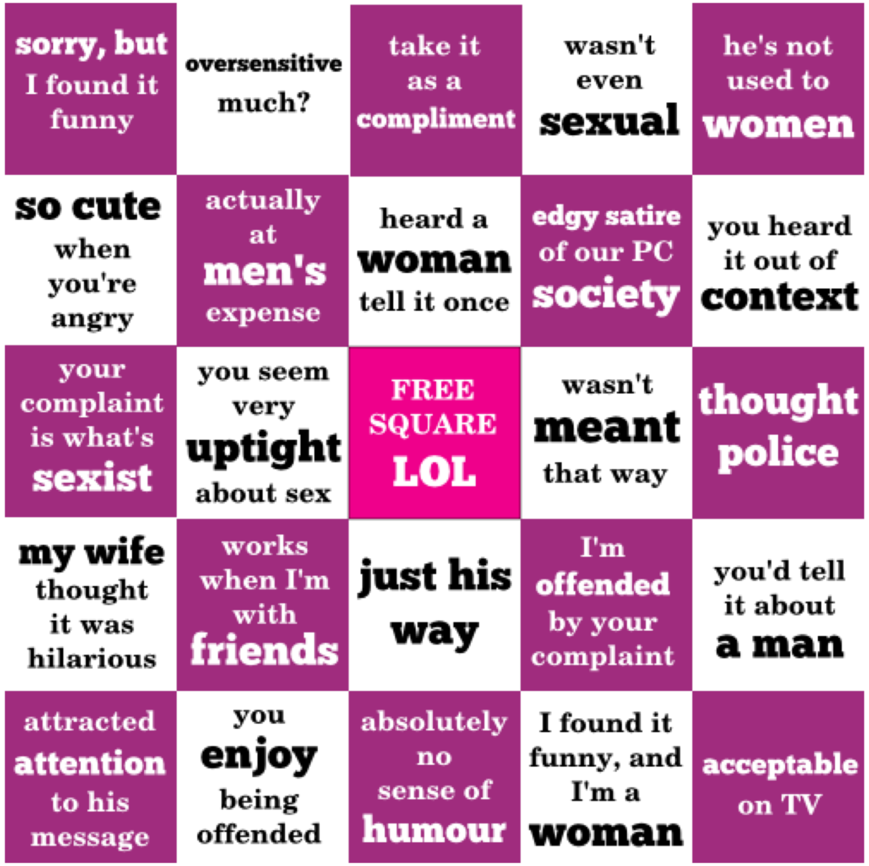
Offensive and discriminatory can enclose many research directions

e.g. hate speech and sexism comments/situations

Krishna et al., A Framework for Cyberbullying Detection in Social Network, 2015
Ying Chen, Yilu Zhou, Sencun Zhu, Heng Xu, oFBI: Detecting Offensive Language in Social Networks for Protection of Youth Online Safety, 2011

Detecting sexism in language

- Over the past decades many different groups have been promoting gender equality
- Woman have gained increasing recognition from the civil society organizations to the United Nations
- **But, this advances not always coincide with changes in society cultural intrinsic behaviors**



http://geekfeminism.wikia.com/wiki/Bingo_card

How can we detect sexism in text?

- The problem can be viewed as follows:
  - Classifying documents (e.g. tweet comment or blog post) as speaking on sensitive topics related to woman
  - Perform sentiment analysis on those documents – objective is to detect negative documents
  - Use these documents to learn more about its language

- Such documents have a high probability of containing profane words, rudeness, sexist jokes, sexual assaults, workplace comments etc.

Joint work from Filipa Peleja and Sara Hajian

By capturing vocabulary used for sexism we can work on answering important questions about sexism in todays society

Sources used for research

Everyday Sexism project

contains textual description about experiences shared by different woman where they were somehow victim of sexism

"I got my first job when I was 16 my former Boss - who was older than twice my age - always made very inappropriate comments. From the first day on he referred to me as "sweetie" instead of my actual name. And though I was employed as part of the office, at the customer support and not his personal secretary he always made me get him his coffee or copying something instead of doing my own work. Soon he started making comments like "I wonder if you take all your orders so good" or "sweetie I can sure show you how to work with that technical stuff." (...)"

6th March 2019

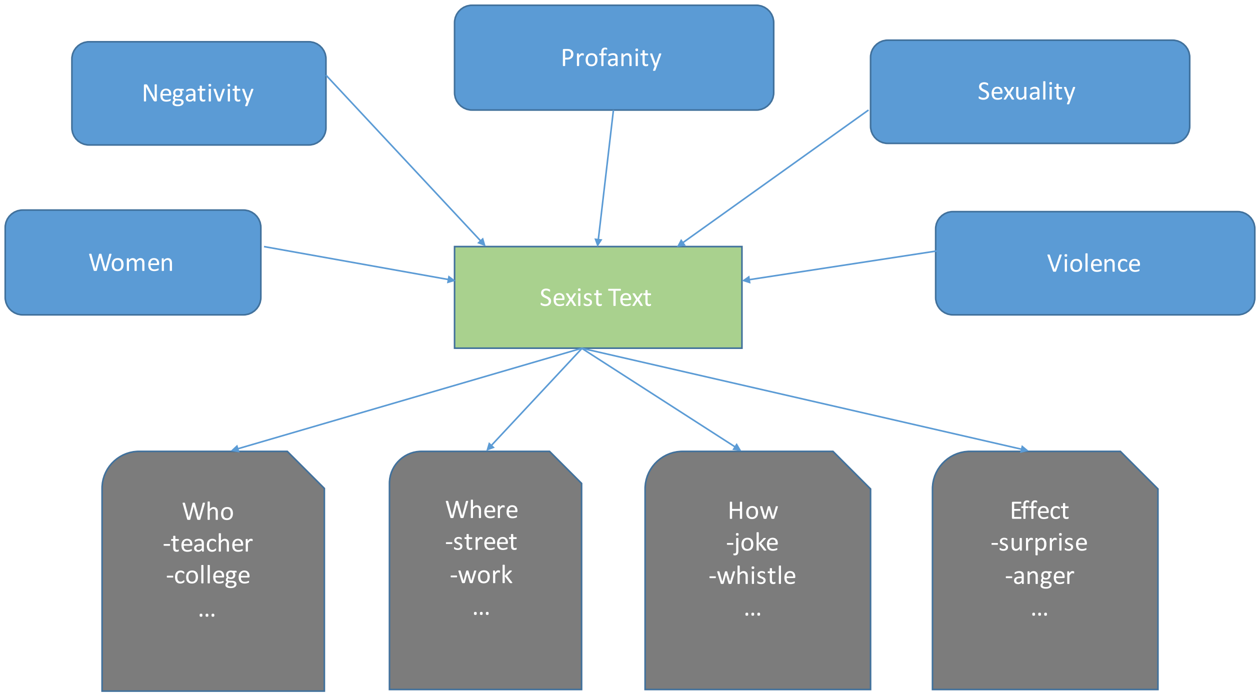
When I was eleven, I was walking into a petrol station after ballet rehearsal. I was wearing my ballet leotard, tights and a cardigan undone as it was a very warm summer. My mum was waiting in the car as I had been desperate for a drink. A couple of men in their late teens or early twenties were standing by their truck filling it up with fuel. One guy yelled "hey gorgeous, wanna f–k?" with the other guy wolf-whistling. Remember that I was eleven years old. I ran into the car and my mum asked me what was wrong. I said the guys scared me and she asked why. I told her that one had yelled a swear word at the other because I didn't 'want to make a fuss'. This is so wrong but I didn't understand at the time what exactly had gone on.

C1

https://everydaysexism.com/

Decomposition of "Everyday Sexism" comments

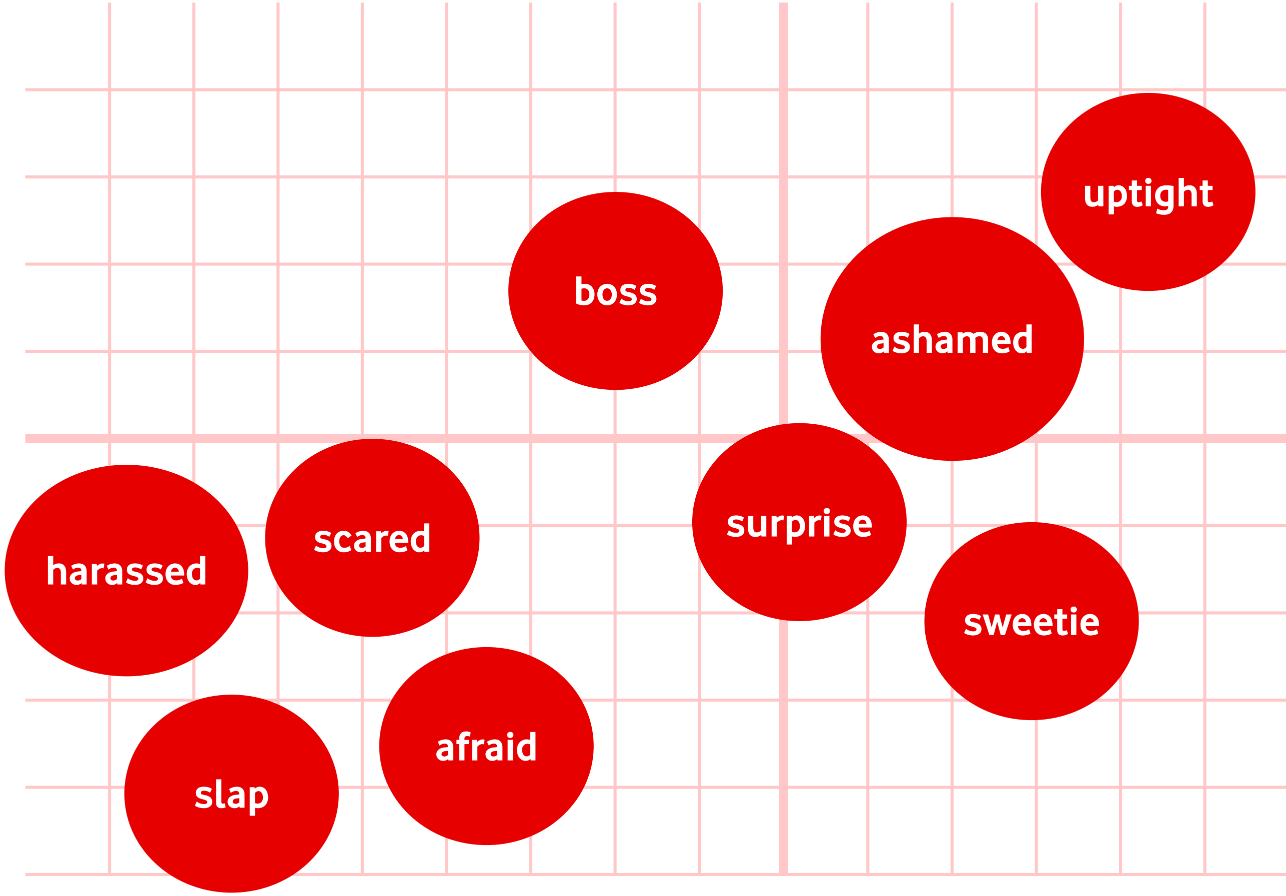
- Sentiment lexicons were used to identify positive and negative tokens
- To identify specific domain vocabulary human annotation was performed



# comments	913
# unigrams	7,153
# bigrams	47,200
# trigrams	93,969
# entities	670
# positive unigrams	167
# negative unigrams	170
# positive bigrams	7,454
# negative bigrams	6,983
# positive trigrams	7,665
# negative trigrams	7,206

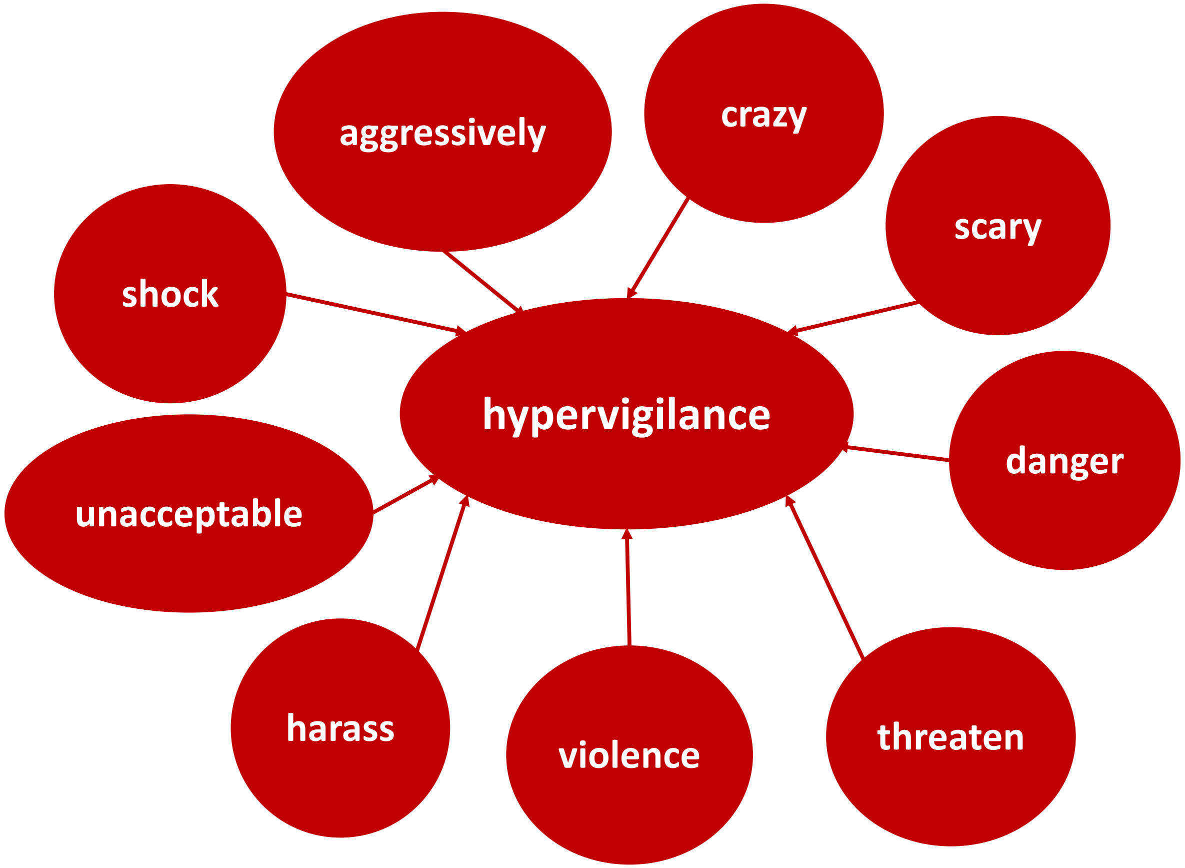
Joint work from Filipa Peleja and Sara Hajian

Word2Vec helps in the task of computing Terms Semantic Similarities



word2vec vectors enclose numerous linguistic regularities and patterns

Most semantically similar terms to the term *hypervigilance*



C1

Joint work from Filipa Peleja and Sara Hajian

Detect Discriminative Terms

$$\boldsymbol{Discriminative}(\boldsymbol{t}) = \prod_{\boldsymbol{d} \in \boldsymbol{D}} \frac{\boldsymbol{sem}(\boldsymbol{t}, \boldsymbol{d}) + \boldsymbol{\mu P}(\boldsymbol{t})}{\sum_{\boldsymbol{t} \in \boldsymbol{V}} \boldsymbol{sem}(\boldsymbol{t}) + \boldsymbol{\mu}}$$

$$\boldsymbol{sem}(\boldsymbol{t}, \boldsymbol{d}) = \frac{\boldsymbol{z}(\boldsymbol{t}) \cdot \boldsymbol{z}(\boldsymbol{d})}{|\boldsymbol{z}(\boldsymbol{t})| \cdot |\boldsymbol{z}(\boldsymbol{d})|} + \boldsymbol{senti}(\boldsymbol{t}, \boldsymbol{d}) \cdot \boldsymbol{cor}(\boldsymbol{t}, \boldsymbol{d})$$

- d is the discriminative term
- $sem(t, d)$ is the semantic similarity between term t and discriminative term d
- $sem(t)$ sums the semantic similarity of a term t and all other discriminative terms
- μ is the average document length
- $P(t)$ is the probability of term t occurring in a given document

- $senti(t, d)$ is the sentiment weight between term t and discriminative term t (this is computed with VADER[1] algorithm)
- $corr(t, d)$ is the correlation between term t and discriminative term t
- $z(\cdot)$ is the function that calculates the semantic vectors using word2vec vectors

Weight terms by how likely they are generated on a model that observes the textual representation of users comments

[1]VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text (by C.J. Hutto and Eric Gilbert)
Joint work from Filipa Peleja and Sara Hajian

Why work on detection discriminatory terms?

Traditional state-of-the-art sentiment lexicons are not able to detect relevant discriminative terms

hypervigilance, unacceptable, objectification_photo, issue_employee

are not found in traditional sentiment lexicons

Models that are able to detect such terms can help in the task of automatically detect discriminatory text from a collection of documents (e.g. tweets or chat bots)

Joint work from Filipa Peleja and Sara Hajian

Are we there yet?

The objective of this work was to compute a vocabulary that is strongly related to sexism and help automatic models to detect sexism **but this is only the beginning**....

who, how, where and **effect**
are important aspects of discriminatory text
that should be addressed

Joint work from Filipa Peleja and Sara Hajian

Vodafone Data and AI

Join us

We want to welcome more brilliant minds like yours to our global team.

https://careers.vodafone.com/vodafone-analytics

See all our jobs in Big Data

Filipa Peleja
filipa.peleja@vodafone.com